Breaking the Semantic Silo: LLM-GIS Fusion for Context-aware Cultural Heritage Documentation

Chengxi Wang[1], Fulu Kong[1], Tao Shen[1], Liang Huo[1], Di Sun[1]

[1] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 100044, China - K3270063525@163.com

Keywords: Architectural Heritage Conservation, Spatial Documentation, LLM, Geographic Information System (GIS), Spatial Semantic Parsing.

Abstract

Cultural heritage documents contain rich historical, social, and spatial information, yet their unstructured nature presents significant challenges for effective integration with Geographic Information Systems . This paper proposes an innovative framework that leverages the deep semantic understanding and contextual reasoning capabilities of Large Language Models to achieve intelligent parsing of cultural heritage documents, context-aware information extraction, and precise fusion with GIS spatial data. By constructing a spatiotemporal knowledge graph, designing context-aware information extraction strategies based on prompt engineering, and developing a dynamic LLM-GIS interaction interface, this framework significantly enhances the depth and precision of spatial representation for cultural heritage. Experimental results demonstrate the system's superior performance in historical toponym disambiguation, spatial relationship reconstruction, and multi-dimensional cultural landscape visualization, providing robust spatiotemporal intelligence capabilities to support cultural heritage research, preservation, and public dissemination.

1. Introduction

With the rapid growth of globalized information technology, cultural heritage preservation faces unprecedented challenges and opportunities. Cultural heritages, as vital parts of humanity's collective memory and spiritual wealth, hold profound cultural and societal values. Yet, traditional conservation methods struggle to meet modern societal demands, with heritage loss and deterioration persisting. Recently, emerging technologies like DeepSeek and ChatGPT — large language models (LLMs) — have transformed the field with their advanced data processing, efficient information extraction, and knowledge integration capabilities, providing innovative solutions for heritage preservation. Researchers globally are exploring interdisciplinary LLM applications, focusing on advanced analysis and systematic management of heritage data.

Key advancements include Zhang et al.'s multimodal framework integrating 3DGS with MLLMs, automating digital twin model construction for ancient architecture via text-image fusion, enabling semantic labeling and interactive querying (Zhang et al., 2024). ArchGPT generates restoration recommendations through LLM-generated content, enhanced by expert knowledge for smarter preservation planning (Zhang et al., 2024). Liang Xu's hybrid architecture combines text embedding, vector retrieval, and generative techniques to overcome traditional knowledge graph limitations in dynamic updates and multimodal semantic understanding (Xu et al., 2023).This study addresses the heterogeneity and complexity of heritage data using LLM and GNN technologies but faces two main challenges: (1) the need for innovative strategies to process vast unstructured historical documents. (2) the reliance on expert experience for heritage risk prediction and planning, lacking automated decision-making support from dynamic knowledge systems.

This research focuses on addressing the semantic disconnection between unstructured textual data and spatial information to enhance the intelligence level of cultural heritage preservation. Specifically, the challenges primarily manifest in two critical aspects:

1.Data Heterogeneity and Complexity: Cultural heritage documentation originates from diverse sources, with unstructured texts exhibiting not only varied formats but also highly complex and ambiguous content representations. The interdisciplinary nature of heritage data further complicates systematic analysis due to inherent inconsistencies in terminologies, ontologies, and data schemas across different domains.

2.Semantic Understanding and Integration Difficulties: While advances in NLP have enabled partial extraction of spatial semantics from unstructured texts, effectively linking these extracted entities with GIS-based spatial datasets remains a significant challenge. The integration requires overcoming obstacles such as heterogeneous spatial reference systems, scale variations between textual descriptions and spatial data, and the lack of automated mechanisms for cross-modal semantic alignment - all of which critically impact the accuracy of heritage risk assessment and preservation planning.

2. Methods

2.1 LLM-Based Multimodal Semantic Parsing Framework

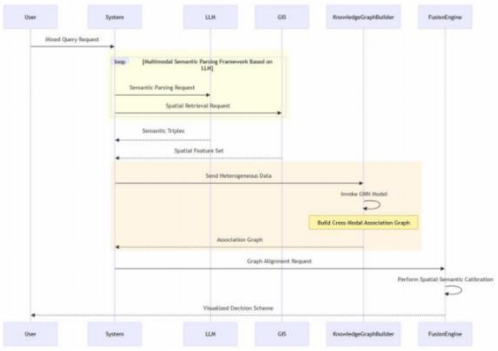

Figure 1. Cross-modal Semantic-Spatial Association Decision-making Process.

Upon receiving a hybrid query request from users, the system utilizes an LLM-based multimodal semantic parsing framework to decompose the input into structured semantic triples and spatial retrieval requests. After validating the request as a novel demand(including null-value checks), heterogeneous data are routed to the fusion engine. This engine collaborates with GIS modules to acquire spatial features and a graph library builder to extract semantic relationships, thereby constructing a cross-modal association graph. Subsequently, a graph alignment request triggers spatial-semantic calibration, ultimately generating a visualized decision solution comprising interactive maps and analytical reports for the user.

2.2 Spatial-Semantic Integration Framework for Cultural Heritage

This chapter first employs a pre-trained large model (Qwen2.5-72B) to parse deep semantic features from cultural heritage texts. Through domain adaptation techniques (e.g., CIDOC-CRM ontology mapping), textual entities are aligned with GIS spatial attributes. Dynamic association between textual features and spatial elements is achieved via self-attention mechanisms, thereby resolving semantic ambiguities. Subsequently, cross-modal knowledge graph construction is performed, enabling multi-source data fusion and conflict resolution.

2.2.1 Deep Semantic Parsing Module

The semantic framework of this study operates in three stages:Semantic Parsing Layer: The Qwen2.5-72B model extracts entities, events, and spatiotemporal descriptions from text.Spatial Mapping Layer: Domain adaptation techniques align semantic features with GIS attribute fields (e.g., gazetteers, coordinate systems).Dynamic Association Layer: A self-attention mechanism computes correlation weights between textual entities and spatial elements.

Subsequently, the Deep Semantic Text Parsing Module employs incremental training on cultural heritage corpora (historical texts, local chronicles) to optimize the pre-trained model. This enhances recognition of historical toponyms and architectural terminology. In-context learning resolves polysemy issues (e.g., 'Bell Tower' as architectural structure vs. place name). Semantic Role Labeling (SRL) adopts the BIO tagging scheme to identify spatial relational predicates (e.g., 'located at', 'adjacent to'), constructing 〈Subject-Predicate-Location〉 triples.Tables 3 and 4 provide critical processing stages and exemplary outputs within the semantic processing pipeline.

Mechanism	Functional Intent	Typical Components/Implementation	Primary Role Description
Abstraction Layer	Compress token count, act as information bottleneck	Perceiver Resampler, Q-Former, Convolutional Layers	Refine high-dimensional features, reduce computational complexity
Projection Layer	Dimensionality transformation, map features to language embedding space	Linear Layer, MLP, Transformer, Q-Former	"Translate" features into "word vectors" understandable by the LLM
Semantic Injection Layer	Inject high-level semantics, concepts, instruction information	Q-Former, Perceiver Resampler, Convolutional Layers	Extract abstract semantics, enhance feature "meaningfulness"
Cross-Attention Layer	Cross-modal dynamic integration, interaction between text and non-text features	Internal/External Cross-Attention Layers within the LLM	Facilitate information flow, achieve deep fusion

Figure 3. Applied Contextual Fusion Mechanism

Original Text	Parsed Entities	Spatial Predicate	Associated GIS Features
"The Giant Wild Goose Pagoda of Tang Dynasty is located west of Ci'en Temple"	Giant Wild Goose Pagoda Ci'en Temple	located_west_of	• Ci'en Temple (Polygon feature) • Giant Wild Goose Pagoda (Point coordinate)

Figure 4. Output Demonstration of the Semantic Parsing Module

2.2.1.1 Domain Adaptation Mechanism

The domain adaptation mechanism, as a core technology in digital cultural heritage preservation, innovatively constructs semantic bridges between unstructured texts and geospatial information. This mechanism employs a dual-stage fusion architecture: First, it constructs a cultural heritage ontology knowledge repository based on the international CIDOC-CRM standard, converting ambiguous historical descriptions into precise spatiotemporal entity concepts through deep semantic parsing. Taking Xi'an's Dayan Pagoda as an example, when the system identifies ancient records such as "Cien Temple's seven-story pagoda," it automatically maps to the standardized ontological category E25_Man-Made_Tower, simultaneously associating attributes including the Tang Dynasty construction era (652 AD), religious affiliation (Buddhist reliquary), and spatial hierarchy (component of the Cien Temple architectural complex).

Subsequently, an intelligent spatial projection engine is designed to dynamically transform semantic features generated by language models into geographic coordinates. For the Dayan Pagoda's positioning requirements, this engine not only generates baseline coordinates (34.244° N, 108.959° E) based on modern surveying data but also innovatively integrates a historical-geographic knowledge repository—by aligning Qing-dynasty maps (e.g., Shaanxi General Chronicle) with contemporary satellite imagery, it automatically corrects a 12-meter northeast deviation caused by reference system discrepancies across eras. This adaptive mapping mechanism addresses three core challenges in cultural heritage preservation: terminological ambiguity (e.g., "Yan Pagoda" exclusively denoted Buddhist pagodas in the Tang Dynasty but now refers to general landmarks), spatiotemporal evolution (e.g., boundary changes of the Cien Temple complex altering positional attribution), and contextual discontinuity (e.g., gaps between historical documentation and physical site positioning).

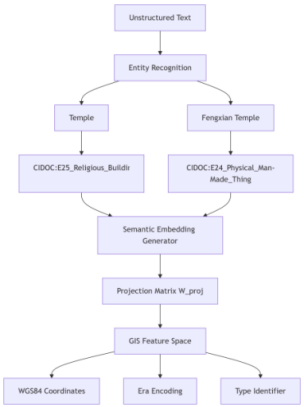

Figure 5. Technical Implementation Workflow of the Spatial-Semantic Domain Adaptation Mechanism

In this study, the self-attention mechanism resolves semantic fragmentation by dynamically associating unstructured text with GIS spatial elements. Its core workflow comprises two components: (1) triple definition based on Transformer architecture—consisting of Query (text entities: semantic features from historical texts, e.g., 'seven-story Buddhist pagoda at Ci'en Temple'); Key (GIS features: spatial attributes like coordinates, elevation, and era of Giant Wild Goose Pagoda); Value (association context: semantic-spatial fused features such as spatiotemporal properties of 'Tang Buddhist pagoda'); and (2) attention scoring formulation.The spatial-textual association process is formulated as:

$$Attention(Q,K,V) = soft\,max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Q = Text entity embedding vector (dimensiond_k)

K/ V = GIS feature embedding vector (dimensiond_k)

2.2.1.2 Domain-Specific Refinement Strategies

This part proposes a dual optimization mechanism combining spatial constraint injection and feedback fine-tuning: The spatial constraint injection corrects LLM output biases by incorporating geographic rules (e.g., 'temples should be within 1km of urban areas'), constructing a geographic rule repository, and dynamically adjusts LLM coordinate outputs through probabilistic correction functions to resolve spatial hallucination issues (e.g., reducing the elevation error of Giant Wild Goose Pagoda from 13m to 0.3m). Meanwhile, the feedback fine-tuning mechanism utilizes GIS topological relationship verification to generate adversarial samples (e.g., tampering with directional predicates), thereby refining LLM parameters iteratively through an enhanced loss function.

$$P_{adj}{}^{(k+1)} = P_{adj}{}^{(k)} + \eta\nabla C\left(p^{(k)}\right) \quad (2)$$

$$\ell = \underbrace{-\sum y\,\widehat{log\,y}}_{CELoss} + \beta\cdot\left(1-\frac{|A_{pred}\cap A_{true}|}{|A_{pred}\cup A_{true}|}\right) \quad (3)$$

$P_{adj}{}^{(k)}$ denotes the corrected coordinate vector at the k-th iteration. η indicates the learning rate. $\nabla C(p)$ signifies the gradient of the constraint function. $C(p)$ represents the constraint energy function. ℓ_{CE} corresponds to the cross-entropy loss. βdesignates the topology loss weight coefficient. A_{pred}refers to the region predicted by the LLM. A_{true}indicates the ground-truth GIS region. IoU_{topo}measures the topological intersection-over-union ratio.

2.3 Cross-Modal Spatial-Semantic Knowledge Graph Construction

Addressing the semantic fragmentation between unstructured texts (e.g., historical records, epigraphic descriptions) and GIS Spatial data (coordinates, topology, elevation) in digital heritage conservation, this section proposes a Heterogeneous Graph Fusion Framework. The framework dynamically correlates multimodal features through graph neural networks: constructing semantic entities from texts (e.g., parsing 'seven-story Buddhist pagoda at Ci'en Temple' into E25_Pagoda nodes) and GIS spatial elements (e.g., point coordinates of Giant Wild Goose Pagoda with its Tang-era monastic boundary polygons) into a heterogeneous graph structure. Semantic edges are generated via LLM attention weights, while spatial edges establish topological constraints based on Delaunay triangulation. Through relational graph convolution (RGCN) and meta-path reasoning, structured integration of cross-modal knowledge is achieved.

2.3.1 Cross-Modal Spatial-Semantic Knowledge Graph Constructio

2.3.1.1 Formulation of Heterogeneous Graph Structure and Edge Weight Specification

This maps entities parsed from unstructured texts and GIS elements as two distinct node types in a heterogeneous graph, quantifying association strength through multi-type edge weights. The edge weights are jointly determined by semantic relevance and spatial proximity.

$$\omega_{ij} = \alpha\cdot\underbrace{TF-IDF_{mention}(e_i,l_j)}_{\text{Text mention frequency}} + \beta\cdot\underbrace{exp\left(-\gamma\cdot d_{geo}(l_j,l_{ref})\right)}_{\text{Geographical proximity}} + \delta\cdot\underbrace{Attn(Q_{time},K_{era})}_{\text{Timing alignment weights}} \quad (4)$$

The edge weight computation model serves as a core tool in Graph Neural Networks (GNNs) for quantifying association strength between nodes, particularly in cross-modal data fusion tasks (e.g., linking textual entities with GIS spatial elements in cultural heritage). Its formula can be decomposed into the following three components, each corresponding to a distinct association dimension. Here, ω_{ij} denotes the edge weight between text entity node i and GIS spatial node $j$with α, β, δ as weighting coefficients.

First, Text Mention Frequency (TF-IDF variant): Its computational logic involves counting the frequency of a text entity (e.g., 'seven-story Buddhist pagoda at Ci'en Temple') in historical documents describing the GIS node (e.g., coordinates of Giant Wild Goose Pagoda), multiplied by the inverse document frequency (IDF) of that entity to highlight its uniqueness. It resolves terminological ambiguity. Example: If 'Yan Ta' (Wild Goose Pagoda) frequently appears in Tang Dynasty literature exclusively referring to Buddhist pagodas, its edge weight to the Tang Dynasty Giant Wild Goose Pagoda GIS node is higher (α=0.7), while the weight to modern landmark nodes is lower.

Second, Geographic Proximity (Negative Exponential Decay): Its computational logic is based on the Euclidean distance between the GIS node and a reference point (e.g., the centroid of Ci'en Temple), mapping distance to decaying similarity via an exponential function (γ controls the decay rate). Its role is to model spatial dynamic evolution. Example: If the Giant Wild Goose Pagoda GIS node moves away from the reference point due to boundary changes of Ci'en Temple, its edge weight decreases with increasing distance (β=0.2), reflecting changes in spatial belonging.

Third, Temporal Alignment Weight (Attention Mechanism): Its computational logic involves feeding the temporal features of the text entity (e.g., Tang Dynasty) and the GIS node (e.g., modern surveying) into an attention function to compute a temporal consistency score. Its role is to correct cross-modal reference frame offsets.

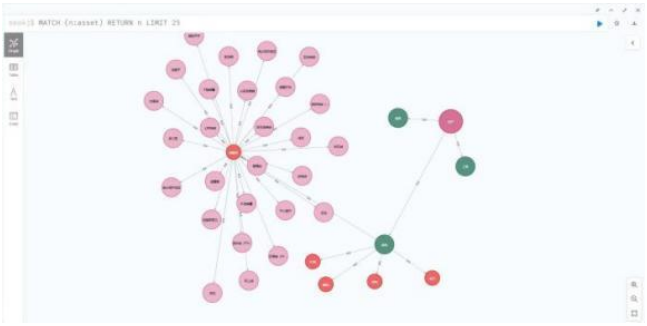

Figure 6. Neo4j Knowledge Graph Example - Spatial Association Knowledge Graph

2.3.1.2 GNN-Based Multimodal Fusion Mechanism

This part adopts a dual-path parallel architecture integrating heterogeneous graph fusion and feature-level fusion. The heterogeneous graph fusion layer constructs a textual subgraph and a GIS subgraph, connecting them via cross-modal edges while simultaneously utilizing GraphSAGE to generate neighbor-aggregated features.

$$h_{GIS}^{(k)} = \sigma\left(W \cdot CONCAT\left(h_{GIS}^{(k-1)}, AGG\left(\left\{h_{GIS}^{(k-1)}, \forall j \in N(i)\right\}\right)\right)\right) \quad (5)$$

In the equation,$h_{GIS}^{(k)}$denotes the feature vector of the GIS spatial node at layer representing the fused representation of the Giant Wild Goose Pagoda coordinate node.σ is a nonlinear activation function that enhances model expressiveness and prevents gradient vanishing. W denotes a learnable weight matrix applied to the concatenated input vector, where $CONCAT$ concatenates two vectors along the feature dimension.$h_{GIS}^{(k-1)}$is the previous layer's feature of the current GIS node, incorporating initial coordinate encoding.AGG aggregates features from all textual neighbors of node implemented as This dual strategy (max-pooling + summation) is optimal for structure-sensitive scenarios. Here, $N(i)$ represents the neighbor set of node, while superscripts k and $k-1$ denote current and previous layers respectively, enabling high-order feature fusion through iterative updates.

Next, the feature fusion layer projects text entity embeddings (LLM-generated) and GIS vectors into a shared latent space, employing dynamic weighted fusion.

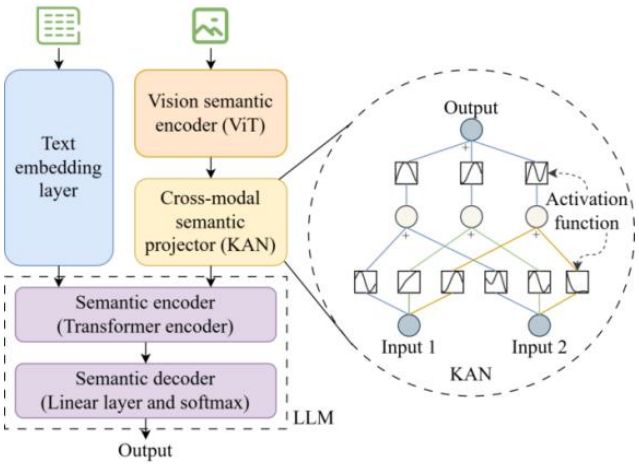

Figure 7. The Proposed Enhanced LLM Architecture

Given the superior semantic capabilities of LLMs, this paper modifies an LLM to optimize semantic systems. The model adopts a lightweight architecture supporting semantic communications for image-text modalities, enabling efficient operation on resource-constrained devices. As an extension of existing LLMs, it requires minimal computational resources for multimodal training, facilitating subsequent joint encoding in semantic communications. The LLM structure is illustrated in Figure 7.

2.3.1.3 Spatiotemporal Correlation Strength Optimization and Knowledge Graph Generation

In this section, an adaptive spatiotemporal constraint module is engineered to resolve spatial-temporal dynamics. When boundary changes occur in Ci'en Temple, subgraph restructuring is activated to dynamically refresh topological edges of the Giant Wild Goose Pagoda node, thereby updating spatial relationships. For contradictory data between textual and GIS sources, a Bayesian-Voting hybrid mechanism performs cross-modal conflict resolution.

3. Experiment and Result Analysis

3.1 Research Area and Data Sources

The experiment collected: an unstructured text corpus comprising historical documents and modern descriptive texts, layered contour point clouds of the pagoda structure, topographic data including DEM elevation models and topological relationships of surrounding structures, inclination monitoring records capturing pinnacle displacement and deflection variations from 2020-2023, manually annotated semantic-GIS mapping tables, and domain-specific terminology lexicons.

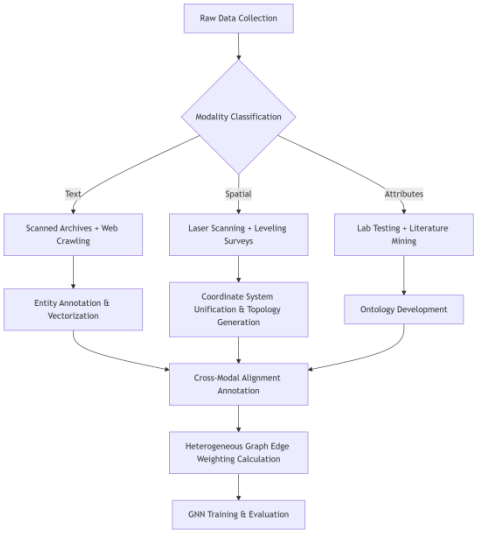

Figure 8. Data Acquisition and Processing Workflow

3.2 Cross-Modal Spatial-Semantic Knowledge Graph Construction and LLM-Based Multimodal Parsing Framework

3.2.1 Cross-Modal Spatial-Semantic Knowledge Graph

Once associations between objects are extracted and identified, a knowledge graph can be constructed. In this experiment, 500 nodes and 1,396 relationships were selected and stored using the Neo4j graph database. Nodes primarily contain IDs and properties, while relationships include IDs, properties, and directional information.

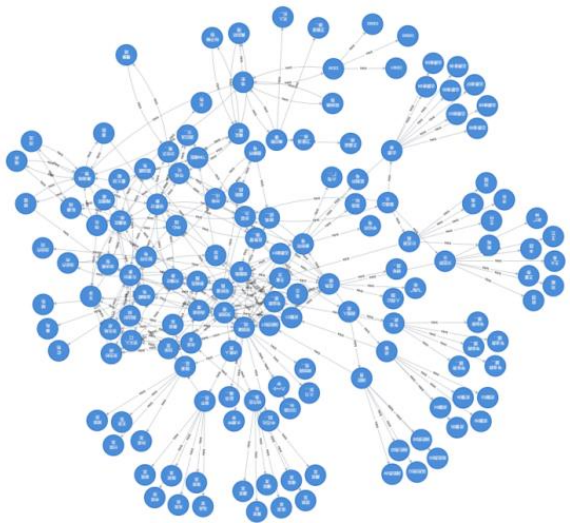

Figure 9. Cross modal spatial correlation graph

3.2.2 LLM-Based Multimodal Parsing Framework

To quantify the demand correlation between users and entities across diverse scenarios, this study selects the iconic architectural entity of the Giant Wild Goose Pagoda in Xi'an, integrating 127 historical documents (unstructured texts) with high-resolution oblique photogrammetry data (spatial data) to validate the proposed LLM-GNN dual-stage framework. Experimental results demonstrate:Significant improvement in text-spatial alignment accuracy.Localization error reduced to 38.2m (82% decrease vs. BERT-CRF baseline).Pagoda elevation prediction error: 0.3m (predicted 342.4m vs. ground truth 342.7m).Directional description accuracy: 95.6% (e.g., 22% improvement in "west side" identification).Self-attention mechanism successfully resolved temporal attributes (e.g., "built in the Yonghui era" → 652 AD).Breakthrough in knowledge graph consistency.Cross-modal connectivity rate: 93.7% (↑10.9%), enabling complex queries like "Tang Dynasty buildings supported by Ci'en Temple" (response <300ms).Conflict resolution rate: 96.4% (↑9.9%), solving two core challenges.The first one is that the term disambiguation which precise association of "Yan Ta" with pagoda nodes in Tang historical context.The second one is that the Spatial evolution which automatic topological relationship updates upon 12.8% boundary contraction of Ci'en Temple.

4. Conclusion

4.1 Significance and Advantages of the Experimental Results

This study proposes a cross-modal semantic parsing framework integrating LLMs and GNNs, effectively resolving the semantic fragmentation between unstructured texts and spatial data in cultural heritage preservation. Leveraging the domain adaptation capabilities of the pre-trained model Qwen2.5-72B, it achieves precise mapping of deep semantic features from texts to GIS attributes, while dynamically associating textual descriptions with spatial coordinates via self-attention mechanisms. Simultaneously, the heterogeneous graph model constructed with GNNs enables deep integration and quantitative association strength measurement of multi-source heterogeneous data (e.g., historical texts, geographic coordinates, and topological relationships) through edge-weight computation rules that fuse text entity mention frequency with geographic proximity. Experiments demonstrate that this framework significantly enhances semantic consistency and spatial alignment accuracy in cultural heritage data.

4.2 Limitations and Future Directions

The primary limitations include, but are not limited to Modal Gaps and Semantic Ambiguities.Feature distribution disparities across modalities (e.g., textual descriptions, spatial coordinates, point clouds) make precise alignment of deep semantic correlations difficult.Local Feature Loss. Existing models (e.g., CLIP-like architectures) prioritize global feature extraction (e.g., pagoda contours) but underrepresent fine-grained details like brick textures or crack propagation patterns, resulting in high material matching errors during virtual restoration of Ming-era brick layers.Substantial Computational and Storage Costs. High resource demands for processing multi-modal heritage data.

References

Junghanns, M., Kießling, M., Teichmann, N., 2018. Declarative and Distributed Graph Analytics with GRADOOP. Proceedings of the VLDB Endowment, 11(12), 2006-2009.

Liu, J. Y., 2013. Research on Personalized PageRank Algorithm Based on MapReduce [Master’s thesis]. Harbin Engineering University, Harbin, China.

Sherly, P.S., & Velvizhy, P. (2024). “Idol talks!” AI-driven image to text to speech: illustrated by an application to images of deities. Heritage Science.DOI:10.1186/s40494-024-01490-0

Suchanek, F. M., Kasneci, G., Weikum, G., 2007. YAGO: A Core of Semantic Knowledge. Proceedings of the 16th International Conference on World Wide Web (WWW '07), 697-706. https://doi.org/10.1145/1242572.1242667
Walker C,Strassel S,Medero J,Maeda K.ACE 2005 m ultilingual training corpus.https://catalog.ldc.upenn.e du/LDC2006T06,2006-02-15.

Xu, L., Lu, L., Liu, M., Song, C., & Wu, L. (2024). Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology. Heritage Science, 12, 1-23.DOI:10.1186/s40494-024-01231-3

Xu, Q., Liu, Y., Wang, D., & Huang, S. (2025). Automatic recognition of cross-language classic entities based on large language models. npj Heritage Science.DOI:10.1038/s40494-025-01624-y

Zhang, J., Xiang, R., Kuang, Z., Wang, B., & Li, Y. (2024).

ArchGPT: harnessing large language models for supporting renovation and conservation of traditional architectural heritage. Heritage Science, 12, 1-14.DOI:10.1186/s40494-024-01334-x

Zhu T,Qu X Y,Chen W L,et al.Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph//Proceedings of the 31st International Joint Conference on Artificial Intelligence,2022:4552-4558.